Investigating Language Independence in HMM PoS/MSD-Tagging

Željko Agić[*], Marko Tadić[**], Zdravko Dovedan[*]

[*]*Department of Information Sciences*
[**]*Department of Linguistics*
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

Abstract. *The paper presents an investigation of functional dependencies in morphosyntactic tagging using hidden Markov models. Starting from a well known fact that the HMM tagging paradigm relies on lexical knowledge acquired from training corpora and stored in form of transition and emission matrices, also called a language model, in the experiment, we apply the TnT trigram tagger on creating language models for seven different languages from MULTEXT East version 3 project translations of George Orwell's novel 1984. – Czech, Estonian, Hungarian, Romanian, Serbian, Slovene and original English version. We then use these language models in the tagging procedure and obtain details on various relations between training corpora statistics, training outputs and outputs of the tagging procedure.*

Keywords. language independence, part-of-speech tagging, morphosyntactic tagging, hidden Markov models

1. Introduction

Hidden Markov models, as described in e.g. [6], are today commonly found in natural language processing tools as the underlying paradigm for part-of-speech/morphosyntactic tagging, being proven as reliable, fast and yielding high tagging accuracy across various languages. A hidden Markov model tagger is usually distributed in form of a computer program implementing two procedures: training and tagging procedure. In course of training, the procedure is fed with previously (manually or otherwise) tagged corpus, creating from it two different matrices of probabilities, called the language model. One matrix represents a probability for a future tag variable to obtain a value t_j when it is already known that previous value is t_j. This matrix is obtained solely by counting overall and specific tag occurrences and is called the transition probability matrix or the n-gram matrix and is usually marked as A. Matrix dimension is governed by the tagging paradigm: a trigram tagger implements a second order HMM, looks two previous tags in order to predict a future tag and makes the matrix three-dimensional. The other matrix captures lexical data, again by counting event occurrences, stating probabilities of symbols (words) being emitted upon reaching states (tags). This matrix is called the emission probability matrix and marked as B. Therefore, language model of a HMM contains two basic knowledge figures: probabilities of tag sequences occurring and probabilities of words being linked to tags. The HMM tagging procedure relies solely on this knowledge as it is – if we set aside specific computational methods used at runtime, such as smoothing procedures and unknown word handlers – the only property of input language known by the tagger.

So if it is known that matrix A of a trigram tagger, containing tag transition counts and probabilities, states all occurrences of tag t_k following t_i and t_j for all tags acquired by the training procedure and if matrix B contains occurrence counts and probabilities for every word-tag pair (w_i,t_j) found by the training procedure, it is safe to claim that overall tagging accuracy depends solely on the training procedure or specifically on (a) number of words acquired on the training corpus and their frequencies and (b) number of acquired sequences of MSD tag uni-, bi- and trigrams and frequencies.

When combined, these two facts also imply another intriguing statement. Being that both matrices are created at training and remain unchanged throughout the tagging procedure – if we ignore a relatively small contribution of

smoothing procedures and other runtime calls of basic HMM taggers to overall accuracy – we are safe to state that tagger accuracy is in fact almost completely defined by the training procedure and not at runtime, being that all HMM taggers use practically the same procedures – Viterbi, linear interpolation, suffix tries, successive abstraction, deleted interpolation – these contributing only by small percentages in implementations of variable quality.

Having stated that overall tagging accuracy is produced mainly by the language model given at training and therefore implicitly defined by the training corpus, i.e. its size (the bigger the corpus, the better the lexical structure image for a language!) and token-MSD distribution, we set off to investigate in detail the various functional dependencies between a corpus as a parameter and tagging accuracies on known and unknown words as output figures. Basically, we consider the training corpus size and all its figures as a set of variables and observe how changes in these variables reflect on tagging accuracy.

In the course of doing so, we had to choose a HMM tagger and training corpora differing in size, lexica and tag sets and to define a proper testing environment. Resources and tools used in the experiment are presented in section 2, the experiment plan is laid out in section 3, while obtained results and discussion can be found in sections 4 and 5 of the paper.

2. Resources and tools

PoS/MSD tagger. For the purposes of this experiment, we chose the well-known TnT trigram tagger [3], implementing the described training procedure, Viterbi algorithm, linear interpolation as a smoothing paradigm and suffix trie, successive abstraction and deleted interpolation as unknown word handling paradigms. In plain words, TnT is a common and commonly used HMM tagger, known for its high speed optimization and language independence and as such – being a *de facto* synonym for HMM MSD-tagging – it was the most reasonable choice.

Besides training and tagging procedures, the TnT software package also contains word counting and accuracy measurement tools. This has enabled us to produce the entire experiment using TnT alone, combined with a small contribution in figures provided by additional helper tools created for purposes of [1].

It should be noted that CroSPoST [2], a newly-developed basic HMM tagger just like TnT or HunPos [5] – currently available as a working beta-version for tagging the Croatian language – could have also been used in this experiment. However, being in beta-version and not yet officially presented to the community nor having an intuitive and fast user interface and accompanying tools, we decided to hold on to TnT for the time being, since our aim was to investigate the general properties of HMM tagging and not to evaluate specific tools used for such investigations.

Corpora. Having said that corpus figures are considered as variables in this experiment, we were obliged to provide a set of corpora that would create a valid testing environment, i.e. produce a significant number of figures. Seven translations of Orwell's novel *Nineteen eighty-four*, obtained from the MULTEXT East version 3 project and MSD tag standard specification [4], seemed to us as a reasonable and valid choice. The translations provided us with the same genre-defined distribution base and language specifics created at translation contributed with intriguing differences in tag subsets, trigram and word distributions. Seven corpora – namely Czech, English, Estonian, Hungarian, Romanian, Serbian and Slovene, omitting Bulgarian because of some character encoding issues that emerged during the preparation stage – were extracted from the specification, preprocessed for TnT and analyzed using helper tools. Basic stats are given in figure 2.1.

	Sentence	Token	Type	MSD
Czech	6752	95828	19117	961
English	6737	**118327**	9769	135
Estonian	6478	90452	17841	410
Hungarian	**6768**	98336	**20318**	405
Romanian	6520	118289	14803	587
Serbian	6677	104313	18094	918
Slovene	6689	107660	17865	**1039**

Figure 2.1 Corpora statistics

We counted sentences, tokens, types and different MSDs assigned to tokens. Highest counts are marked with bold in the figure, giving corpora statistics: Hungarian translation has the most sentences and differing tokens while the original English version of the novel has the highest overall token count and Slovene translation was tagged using the highest number of different MSD tags.

We also provide token counts on adjectives, nouns, pronouns and verbs, proven to be the

most difficult test samples in [1]; details can be found in figure 2.2.

In this figure, part of speech counts are given as first elements of corresponding cells and different MSDs are provided in brackets: e.g. the Czech translation contains 7809 adjectives, tagged by 138 different MSD tags from the MULTEXT East v3 specification.

Some notable differences are actually visible from the table and they might be called language specifics, even from a language independent, statistical point of view: Hungarian language had by far the highest adjective count, noun distribution was similar in all translations and Serbian and Slovene had tagged adjectives and pronouns using a substantial number of MSDs, respectively.

	Adj	Noun	Pro	Verb
Cze	7809 (138)	19293 (80)	11177 (421)	16814 (**148**)
Eng	7426 (4)	21131 (16)	11469 (43)	21348 (29)
Est	5876 (47)	19321 (44)	**12592** (165)	18193 (98)
Hun	**9530** (71)	19972 (153)	6475 (69)	14542 (59)
Rom	7038 (29)	**22688** (34)	11233 (96)	18381 (58)
Ser	7668 (**231**)	20311 (**158**)	9578 (308)	22228 (118)
Slo	7717 (167)	19398 (74)	10861 (**594**)	**25163** (93)

Figure 2.2 PoS and MSD distribution

It could be stated that the choice of corpora does not really contribute to the language independent point of view in this experiment. However, it should be noted that the results obtained and described in following sections can also be verified in [3], [1] and many other experiments with HMM taggers, utilizing corpora differing in both genre (e.g. [1] describes results obtained from a newspaper corpus), tag sets and underlying distributions. Thus, we argue that statistical properties provided by previous figures indeed do set up a valid testing environment with regard to the initial motivation of the experiment. Besides that, the content of the texts in our corpora is also kept under control by using this multilingual parallel corpus. Since we have used an English original and its translations into other languages, we can state that the variations that could be introduced by different content are avoided since this is virtually the same text in different languages. Since all languages will be treated with the same tool, any variation observed could, in fact, be considered the result of language differences.

3. Experiment

Having obtained and prepared the corpora of translations, we had to provide a valid testing framework.

Since the general idea initiating this research plan was detecting general properties of tagging under the HMM paradigm in terms of corpora largely effecting overall tagging accuracy, we chose not to provide a framework that inspects tagging accuracies and then tries to produce reasonable improvement suggestions. Instead, we intended to create such an environment in which we could acknowledge changes in those (overall, known & unknown token) accuracies that reflected changes in figures of training corpora. The framework, however, relies in many ways to the one described in [1], but it does not focus on tagger. It also considers in more detail a flow of these accuracies across languages and their figures.

With regards to dependencies set at section 1, the framework could be subdivided into (a) inspecting overall and known wordform accuracies as a function of lexicon size, i.e. the number of tokens acquired to the emission probability matrix of the language model at training and (b) inspecting overall and unknown wordform accuracies as a function of transition matrix size and quality in the model.

It should also be noted that these two test frames could be subject to inference and pronounced functions of training set size. This is, in fact, the foundation of our framework construction.

Namely, we create nine training sets for each language, with each of these sets S_i containing $i/10$ of the entire corpus. For example, training set S_7 on Slovene would provide 70% of all corpus sentences for the training procedure and the remaining 30% of sentences would be assigned for testing. This assignment scenario ensures us with a fairly large number of unknown words, even in the 90% for training vs. 10% for testing scenario.

The partitioning was also cross-validated for each of the languages in order to ensure a fair testing environment. Therefore, we often state this setup as the worst case scenario, being that it guarantees a large number of unknown words detected by default. Some interesting statistics are provided in figure 3.1 and 3.2.

Note that row values in both figures are given as a function of training set size, left out from the (virtual) first column. It starts at 10% corpus size and moves up to 90% with an iteration step of 10%; hence the nine value rows in these and other figures. Figure 3.1 provides an increase of different token encounters and figure 3.2 presents the same for different MSD tags.

Cze	Eng	Est	Hun	Rom	Ser	Slo
3822	2666	3478	**3891**	3401	3704	3466
6349	4210	5468	**6404**	5524	6048	5942
8618	5268	7799	**8694**	7279	8186	8205
10451	6223	9763	**10622**	8758	10117	9964
12034	6995	11077	**12683**	9824	11549	11519
13961	7632	12586	**14437**	11039	13216	12971
15283	8273	14140	**16048**	12050	14320	14348
16448	8728	15466	**17367**	12983	15833	15598
17844	9309	16637	**18936**	13936	16997	16695

Figure 3.1 Type count and training size

Both functional dependencies are, as expected, logarithmic in nature; the law of logarithm is in a way a governing force of all information science and is by all means expected here. It should also be noted that uni-, bi- and trigram figures would yield the same distribution as tag figures (unigram and tag being the same). However, we omit them for practical purposes – tags are fewer in number when compared to bigrams and trigrams and therefore they fit the tight figures perfectly. Also note that training set sizes are once again omitted from rows in all the figures following these two.

Cze	Eng	Est	Hun	Rom	Ser	Slo
527	113	278	224	275	532	**571**
635	123	324	273	323	646	**680**
703	124	352	304	346	695	**760**
772	128	362	331	350	758	**848**
812	130	375	348	373	784	**871**
849	126	385	364	376	827	**918**
881	132	391	373	383	844	**949**
905	131	394	387	395	878	**989**
927	133	404	395	393	898	**1007**

Figure 3.2 Tag count and training size

The testing procedure proceeds as follows:

a) For each language and each train vs. test partitioning, we create a language model using TnT training procedure. This stage inputs corpora and provides language models, training and testing sets as output.
b) We apply the models on testing sets – i.e. we perform the actual tagging using TnT – and produce accuracy info as output.
c) From these outputs, we assemble token and tag counts, accuracy counts on known and unknown tokens, overall accuracies, properties of n-grams and other facts of interest.
d) We present results in section 4 in a manner reflecting our initial questions: we look into unknown word accuracies as functions of tag and n-gram figures, etc.

Results are presented and discussed in the course of the following two sections.

4. Results

The first result we present is straightforward: overall tagging accuracy as a function of train set size, as presented in figure 4.1.

Cze	Eng	Est	Hun	Rom	Ser	Slo
79.84	**91.31**	83.92	88.19	89.73	76.78	81.83
82.34	**93.49**	87.75	91.38	92.94	79.62	85.03
84.95	**94.60**	89.54	93.21	93.23	82.39	87.12
85.81	**94.80**	91.16	93.94	94.16	83.53	87.34
87.22	**94.96**	91.51	94.32	94.69	84.15	88.74
88.56	**95.61**	91.92	94.23	95.38	85.26	89.29
88.29	**95.93**	92.85	95.04	95.25	85.31	89.28
88.37	**95.47**	93.08	95.27	95.37	86.16	89.72
88.38	**96.28**	92.87	95.27	95.68	86.09	90.49

Figure 4.1 Overall accuracy and training set size

In all the test sets, English yields the highest accuracies, as expected when reviewing type and tag count. This is explained by the fact that English possesses the smallest tag subset and lexicon, thus the quality of transition and emission matrices is the highest in its HMM language model. On the other hand, accuracy on Czech, Serbian and Slovene is poor from similar reasons – large tag sets on highly inflective languages make the matrix quality lower with respect to English.

The following two figures – figure 4.2 and 4.3 – should be considered only when paired with figures 3.1 and 3.2, respectively. Namely, figure 4.2 provides accuracies on known tokens, i.e. tokens previously spotted by the training procedure and therefore known to the tagger at runtime and of course easier to tag. However, in terms of research plans, we desired to setup a link between accuracies on known tokens and numbers of different tokens encountered at training. Therefore, each of the cells in figure 4.2 should be considered as functionally linked with a corresponding cell in figure 3.1.

Similar to this precondition, each of the cells in figure 4.3 is bound by a functional link with a corresponding cell in figure 3.2 in order to provide an insight on tag encounters reflecting in accuracies on unknown words, i.e. those cases in which lexical knowledge of a tagger cannot and does not affect the tagging. A tag is assigned solely by trusting the transition matrix and is therefore governed by tag (that is, unigram, bigram and trigram – see previous notes on their relation) counts.

All three figures – namely 4.1, 4.2 and 4.3 – produce a logarithmic growth of accuracies on all languages, indicating that the functional dependency really is constructive in such a way that it contributes the overall accuracy property in a non-deviant manner.

Cze	Eng	Est	Hun	Rom	Ser	Slo
89.92	94.82	93.54	**97.84**	94.49	87.81	91.16
89.76	95.17	93.95	**98.04**	95.64	87.76	91.27
90.40	96.10	94.31	**98.19**	95.60	89.02	92.05
90.55	95.99	94.89	**98.03**	96.18	89.25	91.72
91.27	96.11	95.07	**98.00**	96.01	89.07	92.26
91.85	96.48	94.79	**98.26**	96.53	89.57	92.40
91.79	96.52	95.44	**97.95**	96.45	89.04	91.91
91.49	96.35	95.29	**98.15**	96.50	89.87	92.19
91.54	96.92	95.09	**98.16**	96.67	89.46	92.86

Figure 4.2 Known accuracy

Therefore, we consider that these figures prove the existence of a valid, strong and constituent functional link between parameters and corresponding values. It should also be noted that highest accuracy on known tokens is achieved on Hungarian, implying that the Hungarian lexicon is the most unambiguous one, i.e. most of the entries have only one tag assigned to them.

Cze	Eng	Est	Hun	Rom	Ser	Slo
57.07	71.13	61.18	64.98	**71.63**	46.05	55.21
60.60	**78.91**	69.95	70.81	78.74	49.65	61.33
64.17	**78.48**	70.50	76.16	77.81	51.52	63.89
65.47	**78.65**	74.63	77.25	78.22	54.58	63.59
67.59	78.57	74.59	77.48	**83.12**	56.29	67.94
69.06	80.90	77.36	75.41	**83.51**	58.33	68.47
67.61	**83.40**	77.76	79.45	82.78	59.36	70.78
69.45	77.48	79.21	79.24	**83.08**	58.31	69.84
67.84	81.24	78.59	78.11	**83.22**	58.72	71.30

Figure 4.3 Unknown accuracy

When considering achieved accuracies on different languages as separate values and engaging in a comparison, one should note that unknown word occurrences differ from one language to another, this being a side-effect of the random-natured testing framework. Figure 4.4 provides the unknown word distribution on test samples across languages to consider accuracy figures in a more correct manner.

Cze	Eng	Est	Hun	Rom	Ser	Slo
30.69	14.80	29.72	29.33	20.83	26.41	25.97
25.44	10.35	**25.84**	24.42	15.97	21.37	20.85
20.81	8.53	20.05	**22.59**	13.29	17.67	17.49
18.87	6.82	18.43	**19.65**	11.25	16.49	15.55
17.09	6.55	17.41	**17.94**	10.20	15.00	14.45
14.45	5.59	16.50	**17.61**	8.82	13.80	13.02
14.48	4.48	14.66	**15.69**	8.78	12.58	12.44
14.12	4.64	13.76	**15.22**	8.39	11.75	11.04
13.32	4.08	13.45	**14.40**	7.39	10.96	10.98

Figure 4.4 Unknown word occurrences

These figures provide a new perspective on previous accuracy figures: lowest unknown word percentages are found on English and Romanian, both yielding highest overall results. High results on unknown tokens can only be considered in previously defined terms and clearly indicate a functional relation.

As a closing word in the results section, it should be noted how even the smallest training sets – the ones containing only 10 percent of corpus sentences – provide a figure of around 30% of unknown tokens encountered at tagger runtime. A bold statement could be derived from this fact: being that unknown words do not occur at tagger operation as often as known tokens, not even when applying the smallest training sets as model constructors, it is fairly reasonable and straightforward to argue how known word accuracies make for a larger share of overall accuracy. Furthermore, it could be stated – as a sole conclusion of this experiment – that the strongest functional dependency, placed upon MSD-tagging accuracy using hidden Markov models, is the one set by the number of tokens encountered during training and known to the tagger at runtime.

5. Conclusions and future work

Having set up the cross-validation framework for testing the differences in application of the same HMM MSD-tagging procedure to seven different languages, we have shown that the HMM trigram taggers, namely TnT, behave the same in all cases. We have also shown that, at least for the trigram MSD-tagging paradigm, the most important and language independent parameter that affects the overall accuracy is the

size of training corpus that ensures the highest number of tokens later known to the tagger.

Future work directions could be spread in several ways:

a) Providing a framework test case for Croatian by including a translation of Orwell's novel into the MULTEXT East v3 corpora set. Corpus preparation is in its final stage at the moment of writing this report.
b) Introducing other corpora – genre-specific, differing in size and lexical properties – to the test, providing a more demanding linguistic environment in order to strengthen the claim of language independence of HMM tagger, as we established it in this experiment.
c) Including CroSPoST [2] – once it reaches its final development stage – and possibly other HMM-based taggers such as HunPos [5] in the experiment, underlying motivation being as in previous direction; providing a high correlation figure on various taggers in the same testing scenario.

6. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776, 130-1300646-0645, 036-1300646-1986 and partially by joint Flemish-Croatian project CADIAL.

7. References

[1] Agić, Ž., Tadić, M. (2006). Evaluating Morphosyntactic Tagging of Croatian Texts. Proceedings of the Fifth International Conference on Language Resources and Evaluation. ELRA, Genoa-Paris 2006.

[2] Agić, Ž., Tadić, M., Dovedan, Z. (2008). Combining part-of-speech tagger and inflectional lexicon for Croatian. Proceedings of the Text-Speech-Dialog Conference, Brno, Czech Republic, in press.

[3] Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington 2000.

[4] Erjavec, T. (2004). Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Fourth International Conference on Language Resources and Evaluation. ELRA, Lisbon-Paris 2004, pp. 1535-1538.

[5] Halácsy, P., Kornai, A., Oravecz, C. (2007). HunPos - an open source trigram tagger. Proceedings of the 45th Annual Meeting of the ACL. Association for Computational Linguistics, Prague, Czech Republic, pp. 209-212.

[6] Rabiner, L. (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989, 77/2, pp. 257-286.